

CLAIMS

1. In a system including a document repository, a method comprising:
 - a) determining, automatically, a level of similarity between at least two of a plurality of discrete elements stored in the document repository; and
 - b) storing data representative of a link between the elements based in-part on the level of similarity.
2. The method of claim 1, wherein the document repository includes documents of at least one type selected from the group comprising a plain text document, a formatted text document, a presentation with discrete pages or slides, a diagram, a spreadsheet, programming code, a semi-structured document database, a text document with mark-up language tags, and a fully structured relational database.
3. The method of claim 1, further comprising:
retrieving a document from the repository;
determining a document type and a physical structure for the document; and
identifying one or more conceptually meaningful segments (elements) within the document based on at least one of the document type and the physical structure.
4. The method of claim 1, further comprising:
displaying the link on a display.
5. The method of claim 1, wherein the document repository includes at least two physical repositories.
6. The method of claim 1, further comprising classifying the plurality documents as belonging to one category of a plurality of predetermined categories, the classification being based on at least one of the group comprising a format for the document, a physical structure for the document, a logical structure for the document, a size of the document, a location where the document is stored, and a content of the document..

7. A method for determining a relationship between documents, the method comprising:

- a) retrieving a plurality of documents from a document repository;
- b) segmenting at least two documents of the plurality of documents into a plurality of conceptually meaningful segments;
- c) determining if a segment of one document is related to a segment of another document; and
- d) storing data representative of the relationship.

8. The method of claim 7, further comprising:

- d) selecting documents from the plurality of documents; and
- e) storing the selected documents in a file store;
wherein the step of segmenting further comprises segmenting at least one of the selected documents into a plurality of conceptually meaningful segments.

9. The method of claim 7, further comprising:

- d) classifying the plurality of documents.

10. The method of claim 9, wherein the document repository is organized in accordance with a directory structure, wherein the step of classifying further comprises classifying the plurality of segments based in-part on the directory structure.

11. The method of claim 9, wherein each document comprises a document name, wherein the step of classifying further comprises classifying the plurality of segments based in part on the document name.

12. The method of claim 9, wherein the step of classifying further comprises classifying the plurality of segments as being a segment type selected from a group comprising requirement, design, code, testing, defects, issues and requests.

13. The method of claim 9, wherein the step of classifying further comprises classifying the plurality of segments based in part on a plurality of classification keywords.
14. The method of claim 7, further comprising comparing the plurality of segments.
15. The method of claim 14, wherein comparing further comprises:
 - a) extracting a plurality of terms from the segments; and
 - b) for each segment, determining the frequency of at least one of the plurality of words within the segment.
16. The method of claim 14, wherein the step of comparing further comprises performing a pair-wise cosine similarity analysis among the plurality of segments.
17. The method of claim 7, wherein the document repository includes documents associated with a software project.
18. A method for analyzing a document, comprising:
 - a) receiving a document, the document including data and a document type, the document type having an associated physical structure;
 - b) determining a logical structure of the document based in part on the data;
 - c) selecting a subset of the data based on at least one of the group including the associated physical structure and the logical structure; and
 - d) storing a document segment, the document segment including the selected subset of the data.
19. The method of claim 18, wherein selecting further comprises using an application programming interface to access the subset of data.
20. A system for determining a relationship between documents, the system comprising:
 - a) a retrieval tool for retrieving a plurality of documents from a document

repository;

- b) a segmentation tool for segmenting at least one document of the plurality of documents into a plurality of conceptually meaningful segments; and
- c) a memory configured to store data representative of a link between at least one segment and one selected from the group comprising the plurality of segments and the plurality of documents.

21. The system of claim 20, further comprising:

- d) a selection tool to select documents from the plurality of documents; and
- e) a file store to store the selected documents;

wherein the segmenting tool is further configured to segment at least one of the selected documents into a plurality of segments.

22. The system of claim 20, further comprising:

- d) a classification tool for classifying the plurality of documents.

23. The system of claim 22, wherein the document repository is organized in accordance with a directory structure, wherein the classification tool is further configured to classify the plurality of documents based in-part on the directory structure.

24. The system of claim 22, wherein each document comprises a document name, wherein the classification tool is further configured to classify the plurality of documents based in part on the document name.

25. The system of claim 22, wherein the classification tool is further configured to classify the plurality of documents as being a document type selected from a group comprising requirement, design, code, testing, defects, issues and requests.

26. The system of claim 22, wherein the classification tool is further configured to classify the plurality of documents based in part on a plurality of classification keywords.

27. The system of claim 20, further comprising a comparison tool for comparing the plurality of segments.
28. The system of claim 27, wherein the comparison tool is further configured to:
 - a) extract a plurality of terms from the segments; and
 - b) for each segment, determine the frequency of at least one of the plurality of terms within the segment.
29. The method of claim 27, wherein the comparison tool is further configured to perform a cosine similarity analysis on the plurality of segments.
30. The system of claim 20, wherein the document repository includes documents associated with a software project.